Dear Thomas,

Thank you for sending the three datasets from Sprocket Central Pty Ltd. The table below shows the summary statistics from the three datasets. Please let us know if there are any misunderstandings.

Table name	No. of records	Distinct Customer IDs
Customer Demographic	4000	4000
Customer Address	20000	3494
Transaction Data	3999	3999

We discovered several quality issues while reviewing the three datasets you supplied us. We have highlighted mitigations and recommendations to prevent future occurrences of poor data quality and improve the accuracy of the data used for decision-making.

- **Extra Customer_ids are found in both the "Transactions table" and "Customer Address table" but not in the "Customer Main (Customer Demographic) Table."**
  Mitigation: Please ensure that the provided tables come from the same time frame. For the analysis, we would be excluding records with the additional id for the model training.
  This is an indication that the data received may not be in sync which each other and it can cause a skew in the analysis if there are missing records in the data.
- **Stale Data Records are found in the "Transaction Table" which is more than 3 months.**
  Mitigation: Assuming the data was collected in December 2017, the relevant transactions would begin in October 2017, hence we filtered down from 20000 to 5079 records.
  Recommendation: Ensure that the transaction table does not contain records for more than 3 months.
  Using stale data can affect the predictions of the model and mismatch the needs of the organization.

- **Missing Data in many fields in the Customer Demography and Transactions Table**
  Mitigation: We replaced the missing values of the according to the distribution of the column. Columns that did not have any impact on the analysis were ignored.
  Recommendation: For the customer demographic data, we suggest that these fields are made "compulsory" in the forms the customer is filling in to capture the data for analysis. Also, ensure that the data pipelines and database are checked thoroughly to prevent missing data for the transaction data.
  For the transactions table, we removed the entire records containing missing values in brand, product_line_product_size, standard cost, and product_first_sold_date since there were no useful data in the entire observations.

- **Inconsistent values for the same attribute and multiple formats for dates in different columns (e.g., New South Wales being represented NSW, Female as "Femal" or "F")**
  Mitigation: We used regular expressions to rename the states to the abbreviations and the Gender to the full form.
  Recommendation: We suggest that a dropdown menu be used to collect this data for consistency.
  This could be caused by either a faulty data pipeline or an open-ended form allowing the customers to freely write their gender or state. Additionally, we changed the date in the product_first_sold_date column of the transaction data to match the date format of the transaction_date column for Consistency.

- **An unusable Column in the Dataset called "default" and a questionable Date of Birth for a Customer**
  Mitigation: Drop the unusable column. Use regular expressions to correct the birth date.
  We require more information about this column to deal with the "default" column appropriately. It contains special characters and a lot of unusable data. We also noticed that a man named Jephthah Bachmann is 179 years old today. We assume this is a mistake and I will change his birth date from 1843 to 1943 to remove a century from his age.

The team would continue with the data cleaning, standardization, and transformation for the model analysis. There would be questions that would arise along the way and all assumptions would be documented. As soon as this exercise is completed, we look forward to spending time with your data SME to address all assumptions and ensure it aligns with Sprocket Central's understanding.

Best Regards,
Shammah Anucha.
Data Analyst.